Automatic dialogue scoring for a second language learning system

Jin-Xia Huang[1], Kyung-Soon Lee[2], Oh-Woog Kwon[3], and Young-Kil Kim[4]

Abstract. This paper presents an automatic dialogue scoring approach for a Dialogue-Based Computer-Assisted Language Learning (DB-CALL) system, which helps users learn language via interactive conversations. The system produces overall feedback according to dialogue scoring to help the learner know which parts should be more focused on. The scoring measures are presented, including task proficiency, grammar accuracy, vocabulary knowledge, and syntactic ability, to assess the user performance during the dialogue. A user evaluation is performed on the automatic dialogue scoring results and the generated feedback to collect the feedback from real learners, and to see if the measures are helpful and proper. A discussion is also held about the difference between the automatic dialogues scoring from essay scoring based on the user evaluation.

Keywords: computer-assisted second-language learning system, dialog-based CALL, automatic dialogue scoring, feedback for L2, automatic essay scoring.

1. Introduction

A DB-CALL system usually provides grammar correction feedback with a grammar checker, and discourse feedback via a semantic checker. We have developed GenieTutor (Kwon, Lee, Kim, & Lee, 2015a; Kwon et al., 2015b), which is a DB-CALL system for English learners in Korea. GenieTutor leads dialogues by asking questions on different topics according to given scenarios, language learners answer questions orally, and the system recognises the speech,

1. Electronics and Telecommunications Research Institute / Chonbuk National University, Daejeon, Korea; hgh@etri.re.kr
2. Chonbuk National University, Jeonju, Korea; selfsolee@chonbuk.ac.kr
3. Electronics and Telecommunications Research Institute, Daejeon, Korea; ohwoog@etri.re.kr
4. Electronics and Telecommunications Research Institute, Daejeon, Korea; kimyk@etri.re.kr

How to cite this article: Huang, J.-X., Lee, K.-S., Kwon, O.-W., & Kim, Y.-K. (2016). Automatic dialogue scoring for a second language learning system . In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 190-195). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.560

evaluates if the answers are semantically proper for given questions, and checks grammatical errors and provides feedback (Lee, Kwon, Kim, & Lee, 2015). The dialogues normally consist of two to four turns, and the system provides semantic and grammar error feedback in each user utterance, deciding if the dialogue can move to the next turn. During the development and user tests, we noticed that users would like to know their overall scoring and level after finishing a whole dialogue.

In this paper, we investigate the dialogue scoring measures for the GenieTutor system. The measures include task proficiency, grammar accuracy, utterance length and complexity, vocabulary level and diversity. Synonyms are also provided as suggestions to improve user vocabulary.

2. Measures for automatic dialogue scoring

Speech scoring has focused on restricted and highly predictable speech, mainly evaluating aspects of speaking related features, including pronunciation, intonation, rhythm, and fluency, such as speaking rate or length and distribution of pauses. For automated scoring of unrestricted spontaneous speech, more speaking content related features are adopted, including grammatical accuracy, syntactic complexity, vocabulary diversity, and spoken discourse structure (Chen & Zechner, 2011; Xie, Evanini, & Zechner, 2012). These speaking content related measures are similar to the essay scoring, because they both intend to assess communicative competence (Attali & Burstein, 2004).

Our DB-CALL system has expected user answers for each system utterance, similar with the restricted speech scoring. However, considering it is a dialogue system, some factors of unrestricted spontaneous speech should also be considered in the scoring.

The first measure investigated for dialogue scoring is dialogue proficiency, indicating how fluently the conversation has been maintained. It consists of task turn pass ratio and user utterance pass ratio, where task turn pass ratio is the ratio of the passed turns out of all task turns. For example, there are 3 turns predefined for the scenario, if the learner passes two turns and gives up in the final turn, than the task turn pass ratio is 66.7%.

User utterance pass ratio is the ratio of the passed user utterances out of all user utterances. For example, for a dialogue with two task turns, the learner finished it with five utterances. Then, there are two task turns, two passed user utterances out

of all five user utterances, then the user utterance pass ratio is 40% while the task turn pass ratio is 100%. If a user utterance passes, the task turn is performed by the semantic correctness module (Kwon et al., 2015a).

The second measure is grammar accuracy. Grammar checks have been performed by grammatical error correction modules in each turn (Lee et al., 2015). What dialogue scoring needs to do is compute the accuracy according to the weighted number of grammar errors, dividing this by the total number of words in all user utterances. This measure is the same with essay scoring (Attali & Burstein, 2004).

The third measure is vocabulary, including vocabulary level and diversity. Vocabulary level has five categories, from primary school level to university level. Vocabulary level estimates the user word level according to the word distributions by dividing the number of user words in *i*th category (*n*uw,i) to the number of user words (*n*uw), and compares it with the vocabulary level of the scenario – a scenario which provides correct references from native speakers. Each category has different a weight *w*i. The vocabulary level would be set to one if the dividing result is higher than one.

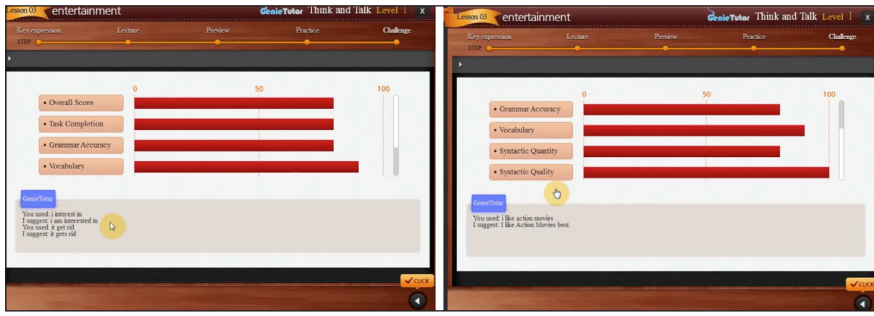
Vocabulary diversity is the ratio of number of word types to tokens in the user utterances (Attali & Burstein, 2004). However, different from essay scoring, the ratio should be a relative one compared with the vocabulary diversity of the scenario. For example, in user utterances "Movies interest me a lot", "I'm interest in Action Movies a lot", there are nine word types, "movies, interest, me, a, lot, I, am, in, action" and 13 tokens, so the diversity is 0.69 (9/13). Again, it is also divided into the vocabulary diversity of the scenario, which would be set to one if the dividing result is higher than one.

The system provides synonyms and similar expressions to improve user vocabulary if the same word is adopted several times in the user utterances. For above cases, GenieTutor will suggest 'See also these similar expressions: interest → fascinate, attract, entertain'.

Syntactic ability includes utterance length and syntactic complexity, the former relates with the utterance lengths, while the complexity considers the syntactic structure of the utterances. Utterance length compares the lengths of user sentences to the length of references in the scenario, and syntactic complexity gives relative complexity scores by considering the length of the utterances and the number of conjunctions.

Dialogue score is the weighted average of all above measures. The system provides overall feedback according to the dialogue score (Figure 1).

Figure 1. Overall feedback of GenieTutor



3. Survey and discussion

A survey on the GenieTutor overall feedback are performed involving 30 human evaluators, 14 of them elementary English learners, and 16 intermediate learners. They are asked five questions, from range one to five, from 'Strongly disagree' to 'Strongly agree', respectively. The evaluators needed to pick at least one item for the final question.

Table 1. User evaluation on the overall feedback generated from dialogue scoring

Idx	Questions		Score
1	Do you think overall feedback would be helpful to improve your conversation level?		3.53
2	Do you think overall feedback would be helpful to motivate your learning?		3.67
3	Do you think the evaluation items of the overall feedback are proper?		3.97
4	Do you think the overall score and the final feedback are proper?		3.50
5	Which evaluation item(s) do you think unnecessary among the overall feedback? You can pick one or more from following items :	A. Overall score	4
		B. Task proficiency	8
		C. Grammar	1
		D. Vocabulary	1
		E. Syntactic	17

From Table 1, we can see that the human evaluators considered the overall feedback *tend to be helpful* to their English learning (average scores=3.53/3.67 for the first and second questions). About the evaluation items of the overall feedback, the users think measuring items are proper (average score=3.97 for the 3rd question). However, the scoring of the items are considered as just *tend*

to be proper (average score=3.50 for the 4th question), implying that the scoring approach still needs to be fine-tuned to reflect the learner's performance more accurately.

Interestingly enough, more than half of the evaluators think the measure *Syntactic* is less necessary (17 votes out of 30 evaluators), while *Vocabulary* and *Grammar* measures get only one vote, respectively. It indicates dialogue scoring should be different from essay scoring considering that the *Syntactic* measure is one of the most important measures in essay scoring. The dialogue in GenieTutor is restricted and predictable, which is very different from the essay. For example, the learner already learns the dialogue "what kind of movies do you like? → I like Action Movies" from the given class (scenario). However when the learner utters the same sentences in the practice, the syntactic complexity measure would give a lower score to the user utterance, and suggest *try to practice longer expressions: I'm interested in Action Movies a lot*. The task proficiency gets eight votes mostly from the speech recognition problem – the learner complains that, when the speech is not recognised correctly, the user utterance would get failure in semantic check, it reduces turn pass ratio and reflects the accuracy of task proficiency. It means the performance of the speech recognition could impede the participants' views.

4. Conclusion

This paper investigated the measures for automatic dialogue scoring and performed user evaluation on the overall feedback. The result showed that the overall feedback after a dialogue tended to be helpful to the language learner, even if there were already turn-by-turn feedback provided for semantic and grammar error correction. The user evaluation result also showed that the dialogue scoring for a DB-CALL system should be different from automatic essay scoring in some measures – in our case, the syntactic ability measure was considered less helpful than others, while grammar and vocabulary measures were considered necessary with the overall score.

5. Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for language learning].

References

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater V.2.0. *Paper presented at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA*. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity fetures for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 722-731).

Kwon, O.-W., Lee, K., Kim, Y.-K., & Lee, Y. (2015a). GenieTutor: a computer assisted second-language learning system based on semantic and grammar correctness evaluations. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 330-335). Dublin Ireland: Research-publishing.net. https://doi.org/10.14705/rpnet.2015.000354

Kwon, O. W., Lee, K., Roh, Y.-H., Huang, J.-X., Choi, S.-K., Kim, Y.-K., Jeon, H. B., Oh, Y. R., Lee, Y.-K., Kang, B. O., Chung, E., Park, J. G., & Lee, Y. (2015b). GenieTutor: a computer assisted second-language learning system based on spoken language understanding. In *Proceedings of the 2015 International Workshop on Spoen Dialogue Systems (IWSDS)*. https://doi.org/10.1007/978-3-319-19291-8_26

Lee, K., Kwon, O.-W., Kim, Y.-K., & Lee, Y. (2015). A hybrid approach for correcting grammatical errors. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 362-367). Dublin Ireland: Research-publishing.net. https://doi.org/10.14705/rpnet.2015.000359

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103-111).

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

CALL communities and culture – short papers from EUROCALL 2016
Edited by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net
Cover design by © Easy Conferences, info@easyconferences.eu, www.easyconferences.eu
Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "bridge" on cover by © Andriy Markov/Shutterstock
Photo "frog" on cover by © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-43-8 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-44-5 (Ebook, PDF, colour)
ISBN13: 978-1-908416-45-2 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2016.